STATISTIQUE EXPLORATOIRE

M1 MASS 2022-2023

PARTIE I : Statistique descriptive univariée et bivariée

Chapitre 1

Les données statistiques

La statistique[1] est une méthode scientifique qui consiste à observer et à étudier une/plusieurs particularité(s) commune(s) chez un groupe de personnes ou de choses.

1.1 Un peu de vocabulaire

Définition 1.1. La *population* est l'ensemble des éléments à étudier ayant des propriétés communes. Un *individu* est un élément de la population étudiée. La *taille* de la population est le nombre d'individus. Un *échantillon* est la partie étudiée de la population.

Exemple. Population : ensemble de parcelles sur lesquelles on mesure un rendement, un groupe d'insectes, élèves d'un groupe de TD, ensemble des accidents d'avion. Individu : une des parcelles, un des insectes, etc.

Remarque. La collecte de données (obtention de l'échantillon à partir de la population) est une étape clé et délicate. Nous ne traitons pas ici des méthodes possibles, mais attirons l'attention sur le fait que l'hypothèse sous-jacente est que l'échantillon d'individus étudiés est choisi "au hasard" parmi tous les individus qui auraient pu être choisis. Il faut tout mettre en œuvre pour que cette hypothèse soit satisfaite. Dans la suite, sauf mention explicite du contraire, nous considérons que l'étude statistique porte sur la population complète.

Définition 1.2. Une *variable* ou *caractère* est une propriété commune aux individus de la population, que l'on souhaite étudier. Elle peut être :

- *qualitative :* lorsque les valeurs prises par la variable ne sont pas une quantité mesurable par un nombre mais appartiennent à un groupe de catégories. On les appelle *modalités* de la variable. On distingue :
  - les variables qualitatives *nominales* : il n'y a pas de hiérarchie entre les différentes modalités ; exemple : sexe, couleur des yeux, couleur de pétales.
  - les variables qualitatives *ordinales* : les différentes modalités peuvent être ordonnées de manière naturelle ; exemple : la mention au baccalauréat, la fréquence d'une activité (jamais, rarement, parfois, souvent, très souvent).

  Remarque. Certaines variables qualitatives peuvent être désignées par un code numérique, qui n'a pas de valeur de quantité. Exemple : le code postal, le sexe (1=garçon, 2=fille).

1. La statistique est à différencier d'"une statistique", qui est un nombre calculé à propos d'une population.

- *quantitative (numérique) :* lorsque les valeurs prises par la variable correspondent à des quantités mesurables et sont données par des nombres. On distingue :
  - les variables quantitatives *discrètes* : elles prennent leurs valeurs dans un ensemble discret, le plus souvent fini ; exemple : le nombre d'enfants, la pointure du pied, le nombre d'espèces recensées sur une parcelle.
  - les variables quantitatives *continues* : elles peuvent prendre toutes les valeurs d'un intervalle réel ; exemple : la taille des individus, le poids d'un individu, le périmètre d'une coquille de moule.

Remarque. L'âge peut être vu et traité comme une variable quantitative discrète ou continue suivant la précision que l'on choisit et le nombre de valeurs qu'il prend au sein de la population. Il peut également exister des variables basées sur l'âge qui sont qualitatives. Si dans un sondage on pose la question "quelle est votre tranche d'âge parmi les possibilités suivantes : - de 25 ans, entre 25 et 40, entre 40 et 60 et + de 60 ans", on peut voir la variable "tranche d'âge" comme une variable qualitative ordinale.

Définition 1.3. L'ensemble des données de la/les variable(s) s'appelle la *série statistique.* Si l'étude statistique porte sur un seul critère, on dit que la série statistique est *simple* (ou *univariée*). Si l'étude porte sur deux ou plusieurs critères, la série est dite respectivement *double* (ou *bivariée*) ou *multiple.*

Exemple. Étudier la longueur des pétales sur une population d'iris donne une série statistique simple ; étudier la longueur et la largeur des pétales donne une série statistique double.

1.2 Notations

Voici les terminologies et notations usuelles pour les définitions ci-dessus.

Terminologie	Notation
Taille de la population	N
Population [2]	$\mathcal{P} = \{1, \dots, N\}$
Individu	$u \in \mathcal{P}$
Variables	$X, Y, \dots$
Donnée de la variable X pour l'individu u	$X(u)$
Série statistique (simple) *brute* pour X	$\{X(1), \dots, X(N)\}$
Série statistique (double) *brute* pour X et Y	$\{(X(1), Y(1)), \dots, (X(N), Y(N))\}$

Table 1.1 – Notations

Exemple 1.1 (Voitures propres). Une petite enquête s'intéresse au constructeur de voiture propre préféré de 11 individus. Les constructeurs proposés sont : Peugeot (P), Renault (R), Citroën (C), Nissan (N), Tesla (T) ; on a la statistique (simple) brute suivante.

Utilisateur	1	2	3	4	5	6	7	8	9	10	11
Constructeur préféré (X)	T	T	R	P	N	C	N	T	P	C	T

2. Il s'avère souvent pratique, voire incontournable (anonymat, etc.), de désigner les individus par des nombres.

La population $\mathcal{P}$ est l'ensemble des 11 individus. La taille de la population est $N = 11$. Les individus sont désignés par des numéros. La variable X étudiée est "le constructeur de voiture propre préféré" ; il s'agit d'une variable qualitative nominale. On a par exemple,

$$X(1) = \text{Tesla}.$$

Exemple 1.2 (Développeuse). Une développeuse d'applications récolte les avis des utilisateurs (de 0 à 5 étoiles). Elle obtient la statistique brute suivante.

Utilisateur	1	2	3	4	5	6	7	8	9	10
Nombre d'étoiles (X)	1	3	5	5	4	2	4	4	5	3

La population $\mathcal{P}$ est l'ensemble des utilisateurs qui ont donné un avis. La taille de la population est $N = 10$. Les utilisateurs sont désignés par des numéros. La variable X étudiée est "l'avis donné" ; il s'agit d'une variable qualitative ordinale. On a par exemple,

$$X(4) = 5 \text{ étoiles}.$$

Exemple 1.3 (Température juillet). On s'intéresse à la température moyenne au mois de juillet dans plusieurs villes de France. On obtient la série statistique brute suivante.

Ville	Température moyenne en juillet (X)
Ajaccio	22.2
Bordeaux	20.8
Clermont-Ferrand	19.7
Brest	16.6
Lille	17.9
Lyon	21.3
Millau	19.3
Nice	23.1
Paris	20
Strasbourg	19.5
Toulouse	21.6
Fort-de-France	27.5
Papeete	25

La population $\mathcal{P}$ est l'ensemble des villes de France considérées. La taille de la population est $N = 12$. Dans ce cas, les individus (les villes) ne sont pas désignés par des nombres, mais par leur nom. La variable X étudiée est "la température moyenne au mois de juillet" ; il s'agit d'une variable quantitative continue. On a par exemple,

$$X(\text{Bordeaux}) = 20,8\,^\circ.$$

1.3 Deux directions en statistique

Il y a deux grandes manières de faire de la statistique : soit descriptive (le sujet de ce cours), soit inférentielle. Nous présentons brièvement les deux approches.

1.3.1 Statistique descriptive

La statistique descriptive a pour but de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses. Quelques questions typiques :

1. Représentation graphique.
2. Paramètres de position, de dispersion, de relation.
3. Régression linéaire.
4. Questions liées à des grands jeux de données (non traité dans ce cours).

1.3.2 Statistique inférentielle

En statistique inférentielle les données ne sont pas considérées comme une information complète, mais une information partielle d'une population infinie. Il est alors naturel de supposer que les données sont des réalisations de variables aléatoires, qui ont une certaine loi de probabilité.

Cette approche nécessite des outils mathématiques plus pointus de théorie des probabilités.

Quelques questions typiques :

1. Estimation de paramètres.
2. Intervalles de confiance.
3. Tests d'hypothèse.
4. Modélisation : exemple (régression linéaire).

La statistique inférentielle n'est pas traitée dans ce cours.

Chapitre 2

Tableaux et représentations graphiques

2.1 Groupement des données

Dans le chapitre précédent nous avons vu des exemples de séries statistiques simples dont les données sont écrites sous forme brute : $\{X(1), \ldots, X(N)\}$. Dans la pratique, le nombre d'individus étant typiquement très grand, il faut réorganiser ces données en les regroupant. La première étape consiste,

- pour une variable qualitative ou quantitative discrète : à identifier les modalités/valeurs prises par la variable, c'est-à-dire à identifier $X(\mathcal{P})$;
- pour une variable quantitative continue : à construire des intervalles ou *classes* formant une partition de l'ensemble des valeurs possibles de la variable. Si possible, on fait en sorte que les classes soient d'amplitude égale, au nombre de 5 à 20 (de préférence entre 6 et 12), contiennent leur borne inférieure mais pas leur borne supérieure (sauf la dernière).

Remarque. Lorsque une variable quantitative discrète prend un grand nombre de valeurs différentes, il est souvent utile de la voir comme une variable quantitative continue et d'effectuer un regroupement en classes. Cela permet une analyse plus claire des données.

Voici les notations utilisées dans ce cours. À noter que dans le tableau ci-dessous on a toujours $p \leq N$.

Terminologie	Notation
Nombre de modalités/valeurs/intervalles pour X	p
Modalités de X, variable qualitative	$X(\mathcal{P}) = \{m_1, \ldots, m_p\}$
Valeurs prises par X, variable quantitative discrète	$X(\mathcal{P}) = \{x_1, \ldots, x_p\}$
Intervalles pour X, variable quantitative continue	$\{[a_0, a_1[, \ldots, [a_{p-1}, a_p]\}$

Table 2.1 – Notations

Exemple. Reprenons les exemples 1.1, 1.2 et 1.3 de la section 1.2.

- *Voitures propres.* On a $p = 5$, et les modalités possibles sont, $m_1 = P, m_2 = R, m_3 = C, m_4 = N, m_5 = T$, autrement dit $X(\mathcal{P}) = \{P, R, C, N, T\}$.
- *Développeuse.* On a $p = 6$, et les modalités possibles sont, $m_1 = 0, \ldots, m_6 = 5$, autrement dit $X(\mathcal{P}) = \{0, \ldots, 5\}$.

- *Température juillet.* La variable étant continue il faut construire des classes. Une solution est de prendre $p = 6$, avec les intervalles :

$$[16, 18[,\ [18, 20[,\ [20, 22[,\ [22, 24[,\ [24, 26[,\ [26, 28[.$$

2.2 Effectifs et fréquences

Afin de retrouver toutes l'information de la série statistique brute en utilisant les regroupements de la section précédente, il faut donner pour chacune des modalités/valeurs/classes son *effectif*. Ceci nous permet ensuite de définir les *fréquences* et *fréquences cumulées*.

Dans toute cette section, nous considérons une population $\mathcal{P}$ de taille N pour laquelle nous étudions une variable X.

2.2.1 Effectifs et fréquences

Rappelons que les modalités/valeurs/classes de la variable X sont notées $m_1,\ \ldots,\ m_p / x_1, \ldots, x_p / [a_0, a_1[, \ldots, [a_{p-1}, a_p]$.

Définition 2.1. Soit $i \in \{1, \ldots, p\}$.

1. Le nombre n_i d'individus pour lesquels la modalité/valeur de la variable X est $m_i/x_i/$dans $[a_{i-1}, a_i[$, est appelé *l'effectif* ou la *fréquence absolue* associé à la modalité/valeur/classe $m_i/x_i/[a_{i-1}, a_i[$:

$$n_i = \text{card}\{u \in \mathcal{P} \,|\, X(u) = m_i / X(u) = x_i / X(u) \in [a_{i-1}, a_i[\}.$$

2. La *fréquence relative* ou simplement *fréquence* associée à la modalité/valeur/classe $m_i/x_i/[a_{i-1}, a_i[$, notée f_i, est le quotient de son effectif par la taille de la population :

$$f_i = \frac{n_i}{N}.$$

Remarque. Pour une variable qualitative ou quantitative discrète, il est équivalent de se donner la série statistique sous forme brute ou sous forme groupée avec les effectifs :

$$\begin{aligned} \text{Variable qualitative :}\ & \{X(1), \ldots, X(N)\} \quad \Leftrightarrow \quad \{(m_1, n_1), \ldots (m_p, n_p)\} \\ \text{Variable quantitative discrète :}\ & \{X(1), \ldots, X(N)\} \quad \Leftrightarrow \quad \{(x_1, n_1), \ldots (x_p, n_p)\}. \end{aligned}$$

Pour une variable quantitative continue, on perd de l'information en regroupant les données car on ne connaît plus la répartition à l'intérieur des classes.

Exemple. Reprenons les exemples de la section 1.2.

- *Voitures propres.* Les effectifs des modalités sont : $n_1 = 2$, $n_2 = 1$, $n_3 = 2$, $n_4 = 2$, $n_5 = 4$.
- *Développeuse.* Les effectifs des modalités sont : $n_1 = 0, \ldots, n_6 = 2$.
- *Température juillet.* Les effectifs des classes sont : $n_1 = 2, \ldots, n_6 = 1$.

Propriété 2.1.

$$\sum_{i=1}^{p} n_i = N, \quad \sum_{i=1}^{p} f_i = 1 \quad \text{et} \quad \forall i \in \{1 \ldots p\},\ 0 \leq f_i \leq 1.$$

Démonstration. La première somme est une autre manière de calculer le nombre total d'individus, soit N. On utilise ensuite la définition de la fréquence pour obtenir

$$\sum_{i=1}^{p} f_i = \sum_{i=1}^{p} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{p} n_i = \frac{N}{N} = 1.$$

L'encadrement de f_i vient du fait que $0 \leq n_i \leq N$. □

A l'aide de ces quantités, on peut construire un tableau qui permet de résumer les données.

Variable X	$m_1/x_1/[a_0, a_1[$	$m_2/x_2/[a_1, a_2[$	$\cdots$	$m_p/x_p/[a_{p-1}, a_p]$	Total
Effectif ou fréquence absolue	n_1	n_2	$\cdots$	n_p	N
Fréquence ou fréquence relative	$f_1 = \frac{n_1}{N}$	$f_2 = \frac{n_2}{N}$	$\cdots$	$f_p = \frac{n_p}{N}$	1

Exemple 2.1 (Pratiques sportives). On étudie la répartition des 9379079 français exerçant une pratique sportive en 2013. Les données sont les suivantes :

Discipline	Aviron	Basket-ball	Cyclisme	Football	Tennis	Autre
Nombre de licenciés	103084	536891	119247	2002398	1111316	5506143

La population $\mathcal{P}$ est l'ensemble des personnes exerçant une activité sportive dans un club en France en 2013. La taille de la population est $N = 9379079$. La variable X est la discipline du sport choisi. Il s'agit d'une variable qualitative nominale dont les modalités sont : aviron, basket-ball, cyclisme, football, tennis et autre.

Pour déterminer la fréquence d'une modalité, il suffit de diviser l'effectif de cette dernière par la taille N de la population. On obtient pour l'aviron par exemple une fréquence égale à $\frac{103084}{9379079} = 0.0109908$. On arrondit ensuite le résultat à quatre chiffres après la virgule afin d'avoir une précision de deux chiffres après la virgule si on exprime le résultat en pourcentage. On obtient le tableau suivant :

Variable X	Aviron	Basket-ball	Cyclisme	Football	Tennis	Autre	Total
Effectif	103084	536891	119247	2002398	1111316	5506143	9379079
Fréquence	0.0110	0.0572	0.0127	0.2135	0.1185	0.5871	1

2.2.2 Effectifs et fréquences cumulé(e)s

Dans cette section, on exclut le cas des *variables qualitatives nominales.* Si la variable X est qualitative ordinale, on suppose que les modalités $m_1, \ldots, m_p$ sont ordonnées suivant l'ordre croissant naturel (ou hiérarchique ascendant) ; si elle est quantitative discrète, on suppose que les valeurs sont classées en ordre croissant : $x_1 < \cdots < x_p$; si elle est quantitative continue, on a un ordre naturel sur les intervalles : $[a_0, a_1[, \ldots, [a_{p-1}, a_p]$.

Définition 2.2. Soit $i \in \{1, \ldots, p\}$.

1. L'*effectif cumulé croissant* (respectivement *effectif cumulé décroissant*) jusqu'à la modalité/valeur/classe $m_i/x_i/[a_{i-1}, a_i[$, noté ν_i (respectivement $\tilde{\nu}_i$), est la somme des effectifs $n_1, \ldots, n_i$ (respectivement $n_i, \ldots, n_p$) :

$$\nu_i = \sum_{j=1}^{i} n_j \quad \text{et} \quad \tilde{\nu}_i = \sum_{j=i}^{p} n_j.$$

2. La *fréquence cumulée croissante* (respectivement *fréquence cumulée décroissante*) jusqu'à la modalité/valeur/classe $m_i/x_i/[a_{i-1}, a_i[$, notée ϕ_i (respectivement $\tilde{\phi}_i$), est la somme des fréquences $f_1, \ldots, f_i$ (respectivement $f_i, \ldots, f_p$) :

$$\phi_i = \sum_{j=1}^{i} f_j \quad \text{et} \quad \tilde{\phi}_i = \sum_{j=i}^{p} f_j.$$

On obtient alors le tableau complété suivant.

Variable X	$m_1/x_1/[a_0, a_1[$	$m_2/x_2/[a_1, a_2[$	$\cdots$	$m_p/x_p/[a_{p-1}, a_p]$	Total
Effectif	n_1	n_2	$\cdots$	n_p	N
Fréquence	$f_1 = \frac{n_1}{N}$	$f_2 = \frac{n_2}{N}$	$\cdots$	$f_p = \frac{n_p}{N}$	1
Fréquence cum. crois.	$\phi_1 = f_1$	$\phi_2 = f_1 + f_2$	$\cdots$	$\phi_p = 1$	pas de sens
Fréquence cum. décrois.	$\tilde{\phi}_1 = 1$	$\tilde{\phi}_2 = f_2 + \ldots + f_p$	$\cdots$	$\tilde{\phi}_p = f_p$	pas de sens

Exemple 2.2 (Ordinateurs). Un responsable du service marketing d'une grande marque d'ordinateurs (portables et de bureau) fait une enquête dans une université pour savoir combien d'ordinateurs (portables ou de bureau) il y a dans le foyer de chaque étudiant. Les données sont les suivantes.

Nombre d'ordinateurs	0	1	2	3	4	5	Total
Nombre d'étudiants	300	876	1235	984	225	154	3774

La population $\mathcal{P}$ est l'ensemble des étudiants de l'université. La taille de la population est $N = 3774$. La variable X est le nombre d'ordinateurs ; il s'agit d'une variable quantitative discrète. On obtient le tableau suivant :

Nombre d'ordinateurs	0	1	2	3	4	5	Total
Effectif	300	876	1235	984	225	154	3774
Fréquence	0.0795	0.2321	0.3272	0.2607	0.0596	0.0408	1
Fréquence cumulée croissante	0.0795	0.3116	0.6388	0.8996	0.9592	1	pas de sens
Fréquence cumulée décroissante	1	0.9205	0.6884	0.3612	0.1004	0.0408	pas de sens

2.2.3 Amplitude et densité de proportion

Dans cette section on se restreint au cas où la variable X est *quantitative continue*, avec classes $[a_0, a_1[, \ldots, [a_{p-1}, a_p]$. On a alors la définition supplémentaire suivante.

Définition 2.3. Soit $i \in \{1, \dots, p\}$.

1. L'*amplitude* de la classe $[a_{i-1}; a_i[$ est $l_i := a_i - a_{i-1}$.
2. La *densité de proportion* de la classe $[a_{i-1}; a_i[$ est $d_i := f_i / l_i$.

Remarque.

1. La densité de proportion permet de comparer les effectifs dans chaque classe en tenant compte de la taille de ces classes (cf. la notion de densité de population en géographie).
2. Dans le cas de classes qui ont toutes la même longueur, il n'est pas nécessaire de calculer la densité de proportion, il est suffisant d'étudier les fréquences relatives ou absolues (qui sont directement proportionnelles à la densité de proportion).

On obtient alors le tableau complété suivant.

Variable X	$[a_0, a_1[$	$[a_1, a_2[$	$\cdots$	$[a_{p-1}, a_p]$	Total
Effectif	n_1	n_2	$\cdots$	n_p	N
Fréquence	$f_1 = \frac{n_1}{N}$	$f_2 = \frac{n_2}{N}$	$\cdots$	$f_p = \frac{n_p}{N}$	1
Fréquence cumulée crois.	$\phi_1 = f_1$	$\phi_2 = f_1 + f_2$	$\cdots$	$\phi_p = 1$	pas de sens
Fréquence cumulée décrois.	$\tilde{\phi}_1 = 1$	$\tilde{\phi}_2 = f_2 + \ldots + f_p$	$\cdots$	$\tilde{\phi}_p = f_p$	pas de sens
Amplitude	$l_1 = a_1 - a_0$	$l_2 = a_2 - a_1$	$\cdots$	$l_p = a_p - a_{p-1}$	pas de sens
Densité de proportion	$d_1 = f_1 / l_1$	$d_2 = f_2 / l_2$	$\cdots$	$d_p = f_p / l_p$	pas de sens

Exemple 2.3 (Moyennes CC). Un professeur obtient les moyennes de contrôle continu suivantes : $8.25, 14.25, 9.5,\ 14, 6.25, 11.75, 10, 10, 17, 14.75, 8, 6, 12.75, 10, 18.75, 11.5, 11.25, 19.25, 3, 13.5, 12.25, 13, 18.5, 15,\ 17,\ 10, 6, 12.25, 4.75, 16, 10.75, 9.75, 15.5, 12, 16.5, 15.25$.

La population $\mathcal{P}$ est l'ensemble des étudiants du groupe de TD. La taille de la population est $N = 36$. La variable X est la moyenne de CC ; il s'agit d'une variable quantitative discrète, mais vu que le nombre de valeurs différentes est grand, on va effectuer une répartition en cinq classes d'amplitude 4. L'amplitude étant constante, on n'a pas besoin de calculer la densité de proportion. On obtient le tableau suivant.

Moyenne CC	$[0, 4[$	$[4, 8[$	$[8, 12[$	$[12, 16[$	$[16, 20]$	Total
Effectif	1	4	12	12	7	36
Fréquence	0.0278	0.1111	0.3333	0.3333	0.1944	1
Fréquence cumulée croissante	0.0278	0.1389	0.4722	0.8056	1	pas de sens
Fréquence cumulée décroissante	1.0000	0.9722	0.8611	0.5278	0.1944	pas de sens

2.3 Représentations graphiques

Pour chaque type de variables, il existe des représentations graphiques qui illustrent les tableaux obtenus dans la section précédente et permettent une bonne lecture des données.

2.3.1 Variables qualitatives

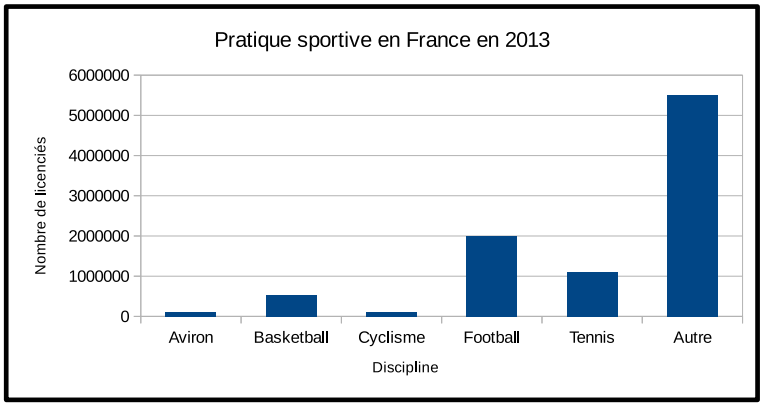
Diagrammes à bandes. Pour obtenir un diagramme à bandes, on trace un repère formé d'un axe horizontal non gradué et d'un axe vertical gradué.

- Sur l'axe horizontal, on trace des rectangles de même largeur, représentant les modalités, que l'on place à des distances régulières les uns des autres
- Les hauteurs des rectangles sont proportionnelles aux effectifs ou aux fréquences des modalités.

Reprenons les données de l'exemple 2.1 (pratiques sportives).

Discipline	Aviron	Basket-ball	Cyclisme	Football	Tennis	Autre
Nombre de licenciés	103084	536891	119247	2002398	1111316	5506143

On obtient le diagramme à bandes suivant.



Remarque.

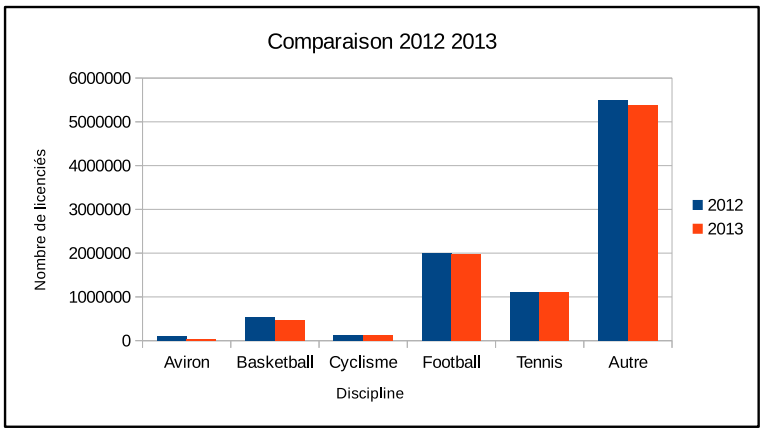
1. Étant donné que les modalités sont qualitatives, il n'y a pas d'ordre entre elles.
2. Ce type de graphique permet aussi de comparer les résultats sur différentes années ou différentes populations.

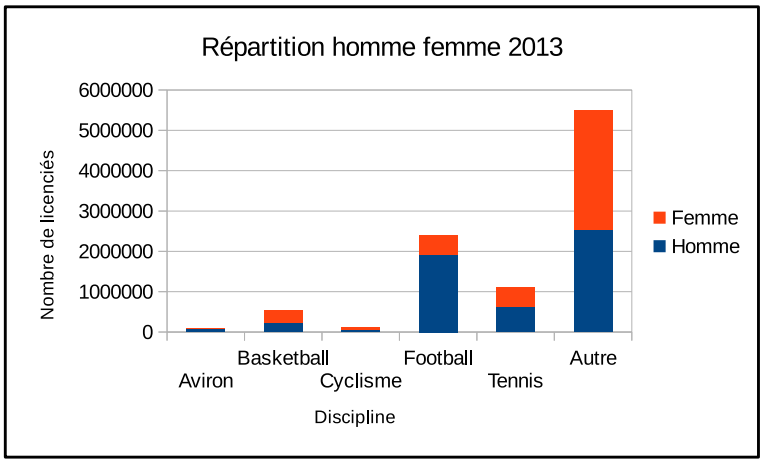
On possède également les données de la répartition des pratiques sportives en France en 2012 :

Année	Aviron	Basket-ball	Cyclisme	Football	Tennis	Autre
2013	103084	536891	119247	2002398	1111316	5506143
2012	43788	475465	115891	1973260	1103519	5369765

On peut aussi donner une représentation de sous-populations. Toujours avec le même exemple, on considère que la répartition entre hommes et femmes est la suivante :

Sexe	Aviron	Basket-ball	Cyclisme	Football	Tennis	Autre
Homme	80810	230532	65312	1912023	629311	2538924
Femme	22274	306359	53935	90375	482005	2967219





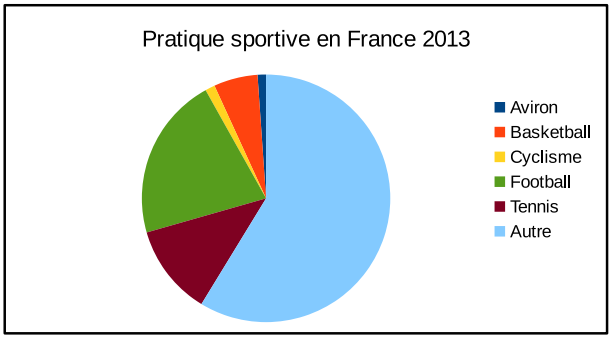
Diagrammes circulaires. Dans le cas d'un diagramme *circulaire* (ou *camembert*), les modalités sont représentées par un secteur angulaire d'un disque (ou d'un demi-disque), dont l'angle est proportionnel à l'effectif ou à la fréquence. On effectue donc un produit en croix pour connaître l'angle de chaque secteur. Dans le cas d'un disque complet, on a ainsi les correspondances suivantes :

$$\begin{aligned} 100\% &\longleftrightarrow 360 \\ x\% &\longleftrightarrow \frac{x \times 360}{100} = x \times 3.6^\circ, \end{aligned}$$

et dans le cas d'un demi-disque, on a :

$$\begin{aligned} 100\% &\longleftrightarrow 180^\circ \\ x\% &\longleftrightarrow \frac{x \times 180}{100} = x \times 1.8^\circ. \end{aligned}$$

Avec l'exemple des pratiques sportives, on obtient :

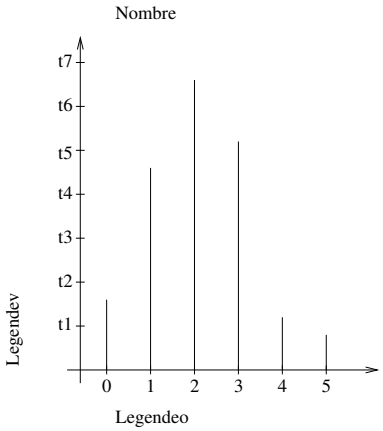


2.3.2 Variables quantitatives discrètes

Diagramme en bâtons. Pour obtenir un diagramme en bâtons, on trace un repère formé de deux axes gradués orthogonaux.

- Sur l'axe des abscisses, on place les valeurs $x_1, \ldots, x_p$ prises par la variable.
- Sur l'axe des ordonnées, on place les effectifs ou les fréquences correspondant aux différentes valeurs.

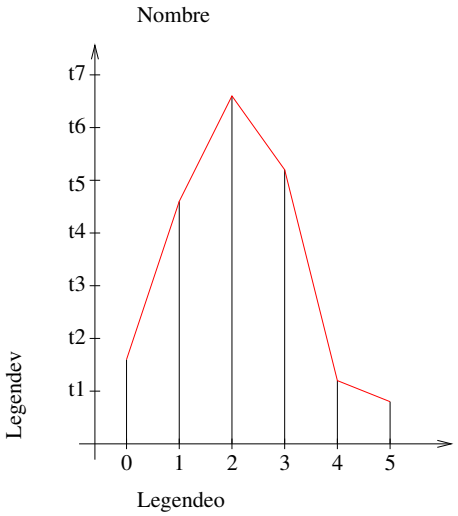
Avec les données de l'exemple 2.2 (ordinateurs), on obtient :



Remarque. Le logiciel Excel ne permet pas vraiment de faire des diagrammes en bâtons. Il faut donc faire un histogramme et réduire au maximum la largeur des tubes.

Définition 2.4. La courbe joignant les sommets des bâtons est appelée *polygone des effectifs*, si l'on a représenté les effectifs, ou *polygone des fréquences*, si l'on a représenté les fréquences.

Avec les données de l'exemple 2.2, on obtient :



2.3.3 Variables quantitatives continues

Histogramme. Pour obtenir un histogramme, on trace un repère formé de deux axes gradués orthogonaux.

- Sur l'axe des abscisses, on place les valeurs des bornes des classes,
- Sur l'axe des ordonnées, on place des pourcentages. Une classe $[a_{i-1}; a_i[$ est représentée par un rectangle de largeur l'amplitude l_i et de hauteur la densité de proportion d_i. Cela revient en fait à ce que l'*aire* de ce rectangle soit la fréquence f_i. En effet, l'aire du rectangle représentant $[a_{i-1}; a_i[$ est :

$$l_i d_i = l_i \frac{f_i}{l_i} = f_i.$$

Voici deux histogrammes pour l'exemple 2.3 (moyennes CC), pour des largeurs de classes différentes. Dans les deux cas, l'aire totale est égale à 1.

X	$[0,4[$	$[4,8[$	$[8,12[$	$[12,16[$	$[16,20]$	Total
Fréquence relative	0.0278	0.1111	0.3333	0.3333	0.1944	1
Densité de proportion	0.0069	0.0278	0.0833	0.0833	0.0486	0.25

X	$[0,8[$	$[8,12[$	$[12,16[$	$[16,20]$	Total
Fréquence relative	0.1389	0.3333	0.3333	0.1944	1
Densité de proportion	0.0174	0.0833	0.0833	0.0486	0.25

Remarque.

1. Il n'est pas possible d'avoir une classe non bornée dans cette représentation car alors l'amplitude est $+\infty$ et la densité de proportion 0. Si une classe est ouverte, il faut *introduire « artificiellement » des bornes finies.*

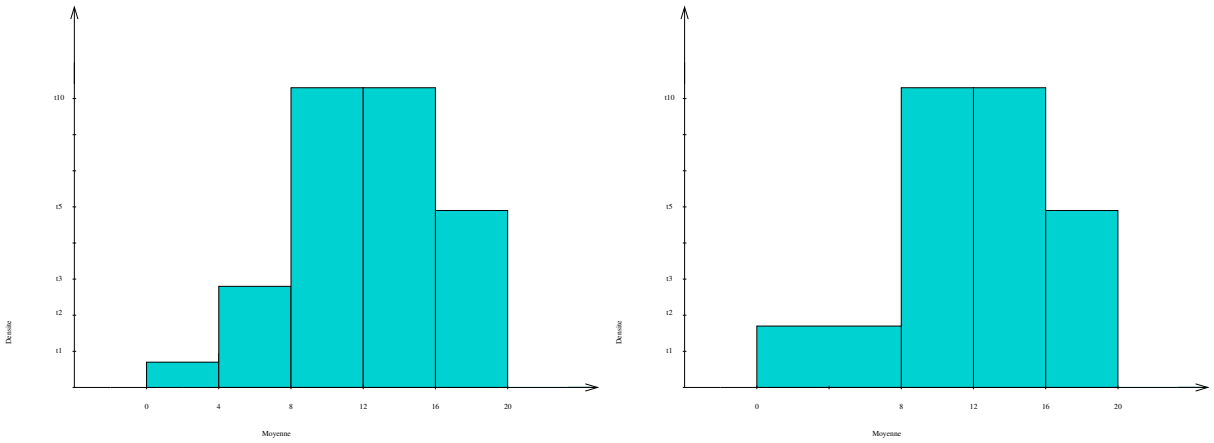
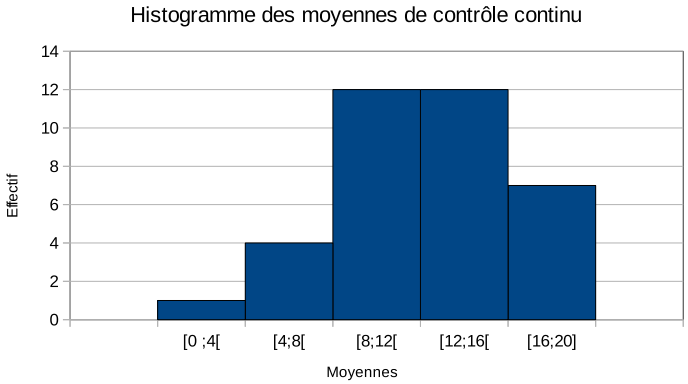


Figure 2.1 – Histogrammes de l'exemple 2.3 avec deux répartitions en classes différentes.

2. Ce qui compte dans l'histogramme est que l'aire d'une barre soit proportionnelle à l'effectif (ou la fréquence) de la classe. Ainsi, si l'amplitude des classes est constante, on peut mettre en ordonnée les effectifs ou les fréquences.
3. Avec un tableur, il ne semble pas possible de tracer un histogramme convenable lorsque l'amplitude n'est pas la même pour toutes les classes. De ce fait, lorsque nous ferons des représentations graphiques avec Excel, nous traiterons des cas où les amplitudes sont constantes.

Voici un histogramme fait avec Excel pour les effectifs des moyennes de contrôle continu, lorsque l'amplitude des classes est constante.



2.4 Fonction de répartition empirique

Dans ce section, nous nous restreignons au cas des variables quantitatives. La fonction de répartition empirique permet de décrire la série statistique de manière complète. La définition diffère légèrement dans le cas discret et continu.

2.4.1 Variables quantitatives discrètes

On suppose que la variable X est discrète et prend comme valeurs $x_1 < \cdots < x_p$.

Définition 2.5. La *fonction de répartition empirique* de X est l'application $F_X : \mathbb{R} \to [0,1]$ construite comme suit :

- $\forall x \in]-\infty, x_1[,\ F_X(x) = 0$;
- $\forall i \in \{1, \ldots, p-1\},\ \forall x \in [x_i, x_{i+1}[,\ F_X(x) = \phi_i$;
- $\forall x \in [x_p, +\infty[,\ F_X(x) = 1$.

Remarque. La fonction F_X est une fonction càdlàg (continue à droite avec limite à gauche).

La fonction de répartition empirique de l'exemple 2.2 (ordinateurs) est donnée ci-dessous.

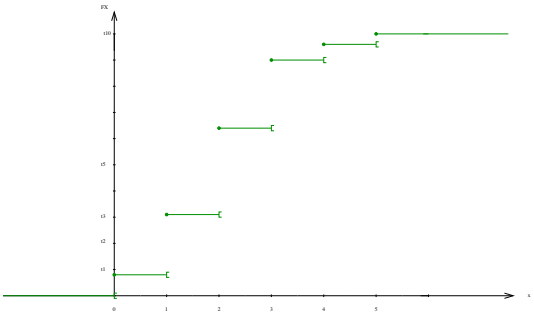


Figure 2.2 – Fonction de répartition empirique de l'exemple 2.2.

2.4.2 Variable quantitative continue

On suppose que la variable X est quantitative continue et a pour classes $[a_0, a_1[, \ldots, [a_{p-1}, a_p]$.

Définition 2.6. La *fonction de répartition empirique* de X est l'application $F_X : \mathbb{R} \longrightarrow [0,1]$ construite comme suit :

- $\forall x \in]-\infty, a_0[,\ F_X(x) = 0$;
- $\forall i \in \{1, \ldots, p\},\ \forall x \in [a_{i-1}, a_i[,\ F_X(x) = \phi_{i-1} + d_i(x - a_{i-1})$;
- $\forall x \in [a_p, +\infty[,\ F_X(x) = 1$.

La fonction de répartition empirique de l'exemple 2.3 (moyennes CC) est donnée ci-dessous.

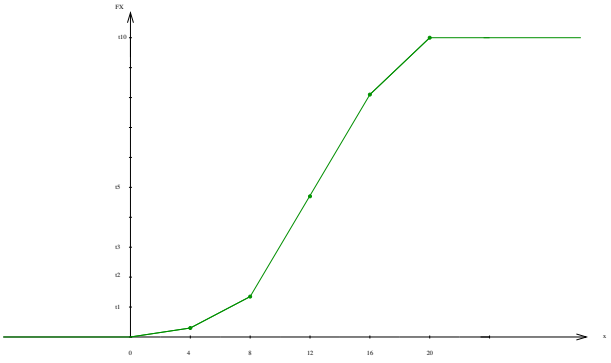


Figure 2.3 – Fonction de répartition empirique de l'exemple 2.3.

Remarque.

1. F_X est définie sur $\mathbb{R}$, affine par morceaux et croissante.
2. Pour tout $i \in \{1, \dots, p\}$, on a :

$$\phi_i = F_X(a_i) \quad \text{et} \quad d_i = \frac{F_X(a_i) - F_X(a_{i-1})}{a_i - a_{i-1}}. \tag{2.1}$$

Proposition 2.1. *L'application F_X est continue sur $\mathbb{R}$.*

Démonstration. L'application F_X est affine par morceaux donc continue par morceaux. Les seuls éventuels points de discontinuité de F_X sont donc les points $a_0, \dots, a_p$ de la subdivision – qui sont également les bornes des classes.

En a_0, on a :

$$\lim_{x \to a_0^-} F_X(x) = \lim_{x \to a_0^-} 0 = 0 \quad \text{et} \quad \lim_{x \to a_0^+} F_X(x) = \lim_{x \to a_0^+} \big(\phi_0 + d_1(x - a_0)\big) = \phi_0 = 0,$$

donc F_X est continue en a_0.

De même, en a_p, on a :

$$\lim_{x \to a_p^-} F_X(x) = \lim_{x \to a_p^-} \big(\phi_{p-1} + d_p(x - a_{p-1})\big) = \phi_{p-1} + d_p(a_p - a_{p-1}) = \phi_p = 1$$
$$\text{et} \quad \lim_{x \to a_p^+} F_X(x) = \lim_{x \to a_p^+} 1 = 1,$$

donc F_X est continue en a_p.

Enfin, fixons $i \in \{1, \dots, p-1\}$. En a_i, on a :

$$\lim_{x \to a_i^-} F_X(x) = \lim_{x \to a_i^-} \big(\phi_{i-1} + d_i(x - a_{i-1})\big) = \phi_{i-1} + d_i(a_i - a_{i-1}) = \phi_i$$
$$\text{et} \quad \lim_{x \to a_i^+} F_X(x) = \lim_{x \to a_i^+} \big(\phi_i + d_{i+1}(x - a_i)\big) = \phi_i,$$

donc F_X est continue en a_i. □

Chapitre 3

Statistique descriptive univariée

Considérons une population $\mathcal{P}$ de taille N, $\mathcal{P} = \{1, \ldots, N\}$, sur laquelle on étudie *une* variable X. Nous nous restreignons dans ce chapitre au cas des variables *quantitative*.

Nous disposons donc d'*une série statistique univariée* à laquelle nous souhaitons associer des paramètres qui synthétisent (réduisent) l'information. Ils sont essentiellement de trois types :

- les paramètres de position,
- les paramètres de dispersion,
- les paramètres de forme.

Rappelons les notations introduites pour la série statistique.

- Les individus sont numérotés de sorte à ce que les valeurs de la variable X soient croissantes. On a alors, la série sous forme brute :
$$X(1) \leq \cdots \leq X(N).$$
- Si la variable X est *quantitative discrète*, on regroupe la série selon les valeurs prises par X en ordre croissant :
$$x_1 < \cdots < x_p.$$
- Si la variable X est *quantitative continue*, on regroupe la série selon les classes de X, naturellement ordonnées :
$$[a_0, a_1[, \ldots, [a_{p-1}, a_p].$$
Pour $i \in \{1, \ldots, p\}$, on note $x_i = \frac{a_{i-1}+a_i}{2}$ le centre de la classe $[a_{i-1}, a_i[$. Rappelons que lors du regroupement en classes, on perd l'information de la répartition à l'intérieure des classes. Dans les calculs de paramètres ci-dessous, on approchera les valeurs de la classes par son centre, et on obtiendra donc des paramètres *approchés* pour les variables continues.

Les exemples que nous donnons dans ce chapitre reprennent les données de l'exemple 2.2 (ordinateurs) pour le cas des variables discrètes. Pour rappel, nous avions obtenu le tableau suivant :

Nombre d'ordinateurs	0	1	2	3	4	5	Total
Effectif	300	876	1235	984	225	154	3774
Fréquence	0.0795	0.2321	0.3272	0.2607	0.0596	0.0408	1
Fréquence cumulée croissante	0.0795	0.3116	0.6388	0.8996	0.9592	1	pas de sens
Fréquence cumulée décroissante	1	0.9205	0.6884	0.3612	0.1004	0.0408	pas de sens

Dans le cas des variables continues, nous utilisons les données de l'exemple 2.3 (moyennes CC) :

Moyenne CC	$[0,4[$	$[4,8[$	$[8,12[$	$[12,16[$	$[16,20]$	Total
Effectif	1	4	12	12	7	36
Fréquence	0.0278	0.1111	0.3333	0.3333	0.1944	1
Fréquence cumulée croissante	0.0278	0.1389	0.4722	0.8056	1	pas de sens
Fréquence cumulée décroissante	1	0.9722	0.8611	0.5278	0.1944	pas de sens

3.1 Paramètres de position

3.1.1 Le mode

Le mode contient l'information des valeurs où les données sont les plus "concentrées".

Définition 3.1.

- Si la variable X est discrète, son *mode* est la ou les valeurs de la variable correspondant à la fréquence maximale.
- Si la variable X est continue, sa *classe modale*, est la ou les classes de densité de proportion maximale.

Exemple.

- *Ordinateurs.* Le mode est égal à 2.
- *Moyennes CC.* Il y a deux classes modales qui sont les classes $[8,12[$ et $[12,16[$.

Remarque. Si X est discrète, une valeur de fréquence maximale est une valeur d'effectif maximal. Cette correspondance *n'a pas lieu* si X est continue car la densité de proportion tient compte de l'amplitude de la classe.

3.1.2 La moyenne (arithmétique)

La moyenne arithmétique représente le *barycentre* de la série statistique, autour de laquelle sont ensuite calculés les paramètres de dispersion.

Définition 3.2.

- Si X est discrète, la *moyenne arithmétique*, notée $\bar{x}$, est la somme des valeurs pondérée par les fréquences :

$$\bar{x} = \sum_{i=1}^{p} f_i x_i = \frac{1}{N} \sum_{i=1}^{p} n_i x_i.$$

  À partir des données brutes, la moyenne arithmétique est le quotient de la somme des valeurs associées à chaque individu par la taille de la population :

$$\bar{x} = \frac{1}{N} \sum_{u=1}^{N} X(u).$$

- Si X est continue, a priori, on ne peut faire qu'un encadrement de la moyenne arithmétique. Cependant, afin de pouvoir calculer un nombre, on appellera *moyenne arithmétique* $\bar{x}$, l'approximation de la moyenne obtenue en prenant comme valeurs les centres des classes :

$$\bar{x} = \sum_{i=1}^{p} f_i x_i = \frac{1}{N} \sum_{i=1}^{p} n_i x_i.$$

Remarque. La moyenne arithmétique a une unité : celle des valeurs de X.

Exemple.

- *Ordinateurs.* Pour la moyenne arithmétique, on obtient :

$$\bar{x} = \frac{1}{3774}\Big(300 \times 0 + 876 \times 1 + 1235 \times 2 + 984 \times 3 + 225 \times 4 + 154 \times 5\Big) = 2.11,$$

ou de manière équivalente :

$$\bar{x} = \Big(0.08 \times 0 + 0.23 \times 1 + 0.33 \times 2 + 0.26 \times 3 + 0.06 \times 4 + 0.04 \times 5\Big) = 2.11.$$

- *Moyennes CC.* La variable étant continue, il faut considérer les centres des classes pour obtenir :

$$\bar{x} = \frac{1}{36}\Big(2 \times 1 + 6 \times 4 + 10 \times 12 + 14 \times 12 + 18 \times 7\Big) = 12.22,$$

ou de manière équivalente :

$$\bar{x} = \Big(2 \times 0.0278 + 6 \times 0.1111 + 10 \times 0.3333 + 14 \times 0.3333 + 18 \times 0.1944\Big) = 12.22,$$

La moyenne possède une propriété intéressante vis-à-vis des changements de variables affines. Précisément :

Proposition 3.1. *Soient $a, b \in \mathbb{R}$ et Y la variable définie par $Y = aX + b$. La moyenne $\bar{y}$ de Y satisfait à l'égalité suivante :*

$$\bar{y} = a\bar{x} + b.$$

Démonstration. Si X est discrète (respectivement continue), alors Y est discrète (respectivement continue) et ses valeurs (respectivement centres des classes) $y_1, \ldots, y_p$ vérifient :

$$\forall i \in \{1, \ldots, p\},\ y_i = ax_i + b.$$

Donc la fréquence de la valeur (respectivement classe de centre) y_i est exactement celle de la valeur (respectivement classe de centre) x_i. Ainsi :

$$\bar{y} = \sum_{i=1}^{p} f_i y_i = \sum_{i=1}^{p} f_i(ax_i + b) = a\Big(\sum_{i=1}^{p} f_i x_i\Big) + b = a\bar{x} + b,$$

comme souhaité. □

3.1.3 Médiane, quartiles, quantiles

Définition 3.3. Supposons que la variable X soit discrète. La *médiane* m est un nombre tel qu'au moins la moitié de l'effectif de la série soit inférieur ou égal à m et au moins la moitié de l'effectif de la série soit supérieur ou égal à m.

- si N est impair, $N = 2P + 1$, la série sous forme brute classée est :

$$X(1) \leq \cdots \leq X(P) \leq X(P+1) \leq X(P+2) \leq \cdots \leq X(2P+1).$$

  La médiane est donc unique et égale à $X(P+1)$.
- si N est pair, $N = 2P$, la série sous forme brute classée est :

$$X(1) \leq \cdots \leq X(P) \leq X(P+1) \leq \cdots \leq X(2P).$$

  Tout nombre de l'intervalle $[X(P), X(P+1)]$ convient. Il existe différentes conventions, telles que $\frac{X(P)+X(P+1)}{2}$ ou $X(P)$.
  Dans le contexte de ce cours, nous fixons la médiane comme étant égale à $X(P)$.

Avec cette convention, la médiane dans le cas où N est pair ou impair satisfait à la définition suivante :

$$m = \min\left\{ x_i \,|\, i \in \{1, \ldots, p\}, \phi_i \geq \frac{1}{2} \right\}.$$

Remarque. La médiane possède une unité : celle des valeurs de X.

Définition 3.4. Supposons que la variable X soit continue. La *classe médiane* est telle que au moins la moitié de l'effectif appartienne aux classes "inférieures ou égales" à cette classe et au moins la moitié appartienne aux classes "supérieures ou égales" à cette classe. De manière analogue au cas discret, il est plus aisé d'avoir une définition qui donne un nombre que l'on puisse calculer. On introduit donc la définition suivante.

La *médiane* d'une variable continue X, notée m, est la *plus petite* solution de l'équation,

$$F_X(x) = \frac{1}{2}, \tag{3.1}$$

où F_X est la fonction de répartition empirique de X. Autrement dit,

$$m = \min\left\{ x \in [a_0, a_p] \,\middle|\, F_X(x) = \frac{1}{2} \right\}.$$

Remarque. Comme X est continue, d'après la proposition 2.1, sa fonction de répartition empirique F_X est continue. Donc F_X vérifie le théorème des valeurs intermédiaires, ce qui montre que l'équation (3.1) admet toujours au moins une solution. Comme l'intervalle $[a_0, a_p]$ est compact, la solution est unique.

Propriété 3.1. *La médiane d'une variable continue X satisfait à l'égalité suivante :*

$$m = a_{i_0-1} + \frac{a_{i_0} - a_{i_0-1}}{\phi_{i_0} - \phi_{i_0-1}} \Big(\frac{1}{2} - \phi_{i_0-1} \Big),$$

où $i_0 = \min\left\{ i \in \{1, \ldots, p\} \middle| F_X(a_i) \geq \frac{1}{2} \right\}$.

Démonstration. La fonction de répartition empirique F_X de X étant croissante, il existe un unique $i_0 \in \{1, \dots, p\}$ tel que :

$$F_X(a_{i_0-1}) < \frac{1}{2} \leq F_X(a_{i_0}), \quad \text{avec} \quad i_0 = \min\Big\{ i \in \{1, \dots, p\} \Big| F_X(a_i) \geq \frac{1}{2} \Big\}.$$

De plus, on sait que $d_{i_0} \neq 0$ car $F_X(a_{i_0-1}) \neq F_X(a_{i_0})$. On résout alors l'équation (3.1) en utilisant l'expression de F_X sur $[a_{i_0-1}, a_{i_0}[$:

$$\frac{1}{2} = F_X(m) = \phi_{i_0-1} + d_{i_0}(m - a_{i_0-1}) = \phi_{i_0-1} + \frac{\phi_{i_0} - \phi_{i_0-1}}{a_{i_0} - a_{i_0-1}}(m - a_{i_0-1})$$

en utilisant les expressions de l'effectif cumulé croissant et de la densité de proportion. On obtient ainsi :

$$m = a_{i_0-1} + \frac{a_{i_0} - a_{i_0-1}}{\phi_{i_0} - \phi_{i_0-1}}\Big(\frac{1}{2} - \phi_{i_0-1}\Big). \qquad \square$$

On peut généraliser le principe de définition de la médiane, ce qui donne lieu à la notion de quantile.

Définition 3.5. Soit $r \in]0, 1]$.

- Si la variable X est discrète, le *quantile d'ordre* r, noté Q_r, est l'unique valeur x_{i_0}, $i_0 \in \{1, \dots, p\}$, telle que $\phi_{i_0-1} < r \leq \phi_{i_0}$:

$$Q_r = \min\{x_i \,|\, i \in \{1, \dots, p\} \text{ et } \phi_i \geq r\}.$$

- Si la variable X est continue, le *quantile d'ordre* r, noté Q_r, est la plus petite solution de l'équation $F_X(x) = r$, où F_X est la fonction de répartition empirique de X :

$$Q_r = \min\{x \in [a_0, a_p] \,\big|\, F_X(x) = r\}.$$

Remarque.

1. Comme la médiane, les quantiles possèdent une unité : celle des valeurs de X.
2. En prenant $r = 1/2$, on retrouve la définition de la médiane, donc $m = Q_{\frac{1}{2}}$.

Définition 3.6.

- Pour tout $n \in \{1, \dots, 99\}$, le n-ième *centile* est $Q_{\frac{n}{100}}$, le quantile d'ordre $n/100$.
- Pour tout $n \in \{1, \dots, 9\}$, le n-ième *décile* est $Q_{\frac{n}{10}}$, le quantile d'ordre $n/10$. Les déciles sont aussi notés $d_1, \dots, d_9$.
- Pour $n \in \{1, 2, 3\}$, le n-ième *quartile*, est $Q_{\frac{n}{4}}$, le quantile d'ordre $n/4$. Les quartiles sont aussi notés q_1, q_2, q_3.

Remarque. La médiane est également le 50-ième centile et le second quartile. De même, les premier et troisième quartiles sont respectivement les 25-ième et 75-ième centiles.

Propriété 3.2. *Le quantile d'ordre r d'une variable continue X satisfait à l'égalité suivante :*

$$Q_r = a_{i_0-1} + \frac{a_{i_0} - a_{i_0-1}}{\phi_{i_0} - \phi_{i_0-1}}(r - \phi_{i_0-1}), \tag{3.2}$$

où $i_0 = \min\{i \in \{1, \dots, p\} \,\big|\, F_X(a_i) \geq r\}$.

Remarque. La médiane est plus robuste que la moyenne : une ou plusieurs données erronées ne font pratiquement pas, voire pas du tout, changer la médiane, alors qu'elles peuvent affecter considérablement la moyenne.

Exemple. Pour le calcul des quantiles, on regarde les fréquences cumulées croissantes.

- *Ordinateurs.* On obtient,

$$m = 2,\ q_1 = 1,\ q_3 = 3,\ d_1 = 1,\ d_9 = 4,\ Q_{\frac{1}{100}} = 0,\ Q_{\frac{99}{100}} = 5.$$

- *Moyennes CC.* On utilise la formule (3.2) pour obtenir :

$$\begin{aligned}
m &= 12 + \frac{4}{0.8056 - 0.4722}\Big(\frac{1}{2} - 0.4722\Big) = 12.34\\
q_1 &= 8 + \frac{4}{0.4722 - 0.1389}\Big(\frac{1}{4} - 0.1389\Big) = 9.33\\
q_3 &= 12 + \frac{4}{0.8056 - 0.4722}\Big(\frac{3}{4} - 0.4722\Big) = 15.34\\
d_1 &= 4 + \frac{4}{0.1389 - 0.0278}\Big(\frac{1}{10} - 0.0278\Big) = 6.60\\
d_9 &= 16 + \frac{4}{1 - 0.8056}\Big(\frac{9}{10} - 0.8056\Big) = 17.94.
\end{aligned}$$

3.1.4 Réflexions sur le mode, la moyenne, et la médiane

Les trois paramètres de *tendance centrale* que sont le mode, la moyenne (arithmétique) et la médiane possèdent un certain nombre de propriétés en commun. Elles sont définies de façon *objective* : à partir des mêmes données, des personnes différentes aboutissent aux mêmes résultats numériques. Elles s'expriment dans la même unité que celle de la valeur de la variable. Elles peuvent être interprétées de façon facile et immédiate. Elles sont simples à calculer.

La moyenne (arithmétique) possède une propriété supplémentaire importante : sa valeur dépend de toutes les observations retenues (si on fait varier une des valeurs observées, la moyenne sera toujours modifiée au contraire du mode et de la médiane qui peuvent rester inchangées). C'est essentiellement pour cette raison que la moyenne, qui est plus longue à calculer que le mode ou la médiane, est considérée comme le principal paramètre de tendance centrale.

3.2 Paramètres de dispersion

3.2.1 Étendue et distance interquartile

Rappelons qu'avec nos conventions, la plus petite valeur de la série est $X(1)$ et la plus grande est $X(N)$.

Définition 3.8. L'*étendue* de la série, notée δ_e, est la longueur du segment $[X(1), X(N)]$:

$$\delta_e = X(N) - X(1).$$

Remarque.

1. L'étendue possède une unité : celle des valeurs de X.
2. L'étendue *n'est pas très informative* car elle ne tient pas compte de la répartition des données dans le segment $[X(1), X(N)]$.

Définition 3.9. L'*intervalle inter-quartile* est le segment $[q_1, q_3]$. Il contient la moitié « centrale » des données de la série. La *distance inter-quartile*, notée δ_q, est la longueur de l'intervalle inter-quartile :

$$\delta_q = q_3 - q_1.$$

Remarque.

1. La distance inter-quartile possède une unité : celle des valeurs de X.
2. La distance inter-quartile est plus précise que l'étendue car elle ne tient compte que de des données « proches » de la médiane en un certain sens.

3.2.2 Variance et écart-type

Définition 3.10. Si X est discrète (respectivement continue), sa *variance*, notée $\mathrm{Var}(X)$, est la moyenne des écarts quadratiques des valeurs (respectivement des centres des classes) de X à sa moyenne :

$$\mathrm{Var}(X) = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2.$$

Remarque.

1. La variance *n'a pas* la même unité que les valeurs de X, mais celle de leur carré.
2. Lorsque la variable est discrète, avec les données brutes, la variance de X est le quotient de la somme des distances quadratiques à la moyenne des valeurs associées à chaque individu par la taille de la population :

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{u=1}^{N}\big(X(u) - \bar{x}\big)^2.$$

Proposition 3.2. *La variance de X s'exprime également comme :*

$$\mathrm{Var}(X) = \left(\sum_{i=1}^{p} f_i x_i^2\right) - \bar{x}^2 = \Big(\frac{1}{N}\sum_{i=1}^{p} n_i x_i^2\Big) - \bar{x}^2 = \Big(\frac{1}{N}\sum_{u=1}^{N} X(u)^2\Big) - \bar{x}^2.$$

Démonstration. Montrons la première égalité, les deux autres étant simplement des réécritures. Il s'agit d'utiliser une identité remarquable :

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{i=1}^{p} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\\ &= \sum_{i=1}^{p} f_i x_i^2 - 2\bar{x}\sum_{i=1}^{p} f_i x_i + \bar{x}^2\sum_{i=1}^{p} f_i\\ &= \sum_{i=1}^{p} f_i x_i^2 - 2\bar{x}(\bar{x}) + \bar{x}^2(1),\end{aligned}$$

par application de la définition de la moyenne $\bar{x}$ et de la propriété 2.1. Le résultat suit. □

Exemple 3.1.

- *Ordinateurs.* On utilise la proposition précédente :

$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{3774}\Big(300\times 0^2 + 876\times 1^2 + 1235\times 2^2 + 984\times 3^2 + 225\times 4^2 + 154\times 5^2\Big) - 2.11^2\\ &= 1.40\end{aligned}$$

  ou de manière équivalente :

$$\begin{aligned}\mathrm{Var}(X) &= \Big(0.08\times 0^2 + 0.23\times 1^2 + 0.33\times 2^2 + 0.26\times 3^2 + 0.06\times 4^2 + 0.04\times 5^2\Big) - 2.11^2\\ &= 1.40.\end{aligned}$$

- *Moyennes CC.* La variable étant continue, il faut considérer les centres des classes pour obtenir :

$$\mathrm{Var}(X) = \frac{1}{36}\Big(2^2\times 1 + 6^2\times 4 + 10^2\times 12 + 14^2\times 12 + 18^2\times 7\Big) - 12.22^2 = 16.39,$$

ou de manière équivalente :

$$\begin{aligned}\mathrm{Var}(X) &= \Big(2^2\times 0.0278 + 6^2\times 0.1111 + 10^2\times 0.3333 + 14^2\times 0.3333 + 18^2\times 0.1944\Big) - 12.22^2\\ &= 16.39,\end{aligned}$$

Comme somme de carrés, la variance est un nombre positif. C'est un bon indicateur de dispersion des données car la variance s'annule si et seulement si toutes les observations effectuées sont :
— identiques si X est discrète ;
— dans la même classe si X est continue.
Précisément, on a le résultat suivant :

Théorème 3.1. *La variance de X s'annule si et seulement si :*

$$\exists\, i_0 \in \{1, \dots, p\},\ \forall i \neq i_0,\ n_i = 0. \tag{3.3}$$

Auquel cas, un tel i_0 est unique, donc toutes les observations sont identiques (respectivement dans la même classe) si X est discrète (respectivement continue).

Démonstration. Supposons l'assertion (3.3) vérifiée. Alors $N = n_1 + \cdots + n_p = n_{i_0}$ et la moyenne de X est :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \frac{1}{n_{i_0}} n_{i_0} x_{i_0} = x_{i_0},$$

et la variance de X vaut :

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{p} n_i (x - \bar{x})^2 = \frac{1}{n_{i_0}} n_{i_0} (x_{i_0} - x_{i_0})^2 = 0.$$

Réciproquement, supposons $\mathrm{Var}(X) = 0$,

$$\mathrm{Var}(X) = 0 \iff \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = 0 \iff \sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = 0.$$

$$\iff \forall i \in \{1, \dots, p\},\ n_i (x_i - \bar{x})^2 = 0,$$

car une somme de termes positifs n'est nulle que si chacun des termes la composant est nul. Ainsi, on obtient :

$$\forall i \in \{1, \dots, p\},\ \big(n_i = 0 \text{ ou } x_i = \bar{x}\big).$$

Comme $N \neq 0$, il existe $i_0 \in \{1, \dots, p\}$ tel que $n_{i_0} \neq 0$, ce qui entraîne $\bar{x} = x_{i_0}$. Soit alors $i \neq i_0$. On a $x_i \neq x_{i_0}$, donc $x_i \neq \bar{x}$, ce qui entraîne $n_i = 0$ et conclut la preuve. □

Définition 3.11. L'*écart-type* de X, noté σ_X, est la racine carrée de la variance : $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

Remarque. L'écart-type est également un bon indicateur de dispersion car il vérifie la propriété du théorème 3.1. Mais de plus, l'écart-type a *la même* unité que les valeurs de X.

Exemple. Grâce aux calculs de l'exemple 3.1, on obtient directement $\sigma_X = 1.18$ pour le nombre d'ordinateur et $\sigma_X = 4.05$ pour les notes.

De même que pour la moyenne à la proposition 3.1, on peut étudier le comportement de la variance et de l'écart-type par changement de variables affine.

Théorème 3.2. *Soient $a, b \in \mathbb{R}$ et Y la variable définie par $Y = aX + b$. Alors :*

$$\mathrm{Var}(Y) = a^2 \,\mathrm{Var}(X) \quad \textit{et} \quad \sigma_Y = |a|\sigma_X.$$

Démonstration. Si la variance se comporte comme annoncé, on obtient :

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \sqrt{a^2\,\mathrm{Var}(X)} = |a|\sqrt{\mathrm{Var}(X)} = |a|\sigma_X.$$

Reste donc à montrer l'égalité sur les variances. Pour ce faire, on écrit :

$$\mathrm{Var}(Y) = \sum_{i=1}^{p} f_i(y_i - \bar{y})^2,$$

avec f_i fréquence de la valeur (respectivement classe de centre) x_i si X est discrète (respectivement continue) et $y_i = ax_i + b$ (voir proposition 3.1). Donc :

$$\mathrm{Var}(Y) = \sum_{i=1}^{p} f_i\big((ax_i + b) - (a\bar{x} + b)\big)^2 = a^2 \sum_{i=1}^{p} f_i(x_i - \bar{x})^2 = a^2\,\mathrm{Var}(X),$$

comme escompté. □

3.3 Box plot ou boîte à moustaches

La boîte à moustaches, ou "box plot" en anglais, est une représentation graphique mettant en exergue certains des paramètres de position et de dispersion précédents. Elle permet de visualiser sommairement la répartition des données de la série statistique et les données "extrêmes". Elle est constituée :

- d'un diagramme en boîte : un rectangle de bornes q_1, q_3 coupé au niveau de la médiane m et éventuellement également au niveau de la moyenne $\bar{x}$ (alors signifié en pointillés) ;
- de moustaches : deux segments joignant q_1 à $X(1)$ pour l'un et q_3 à $X(N)$ pour l'autre.

La figure 3.1 représente l'allure générale d'une telle représentation.

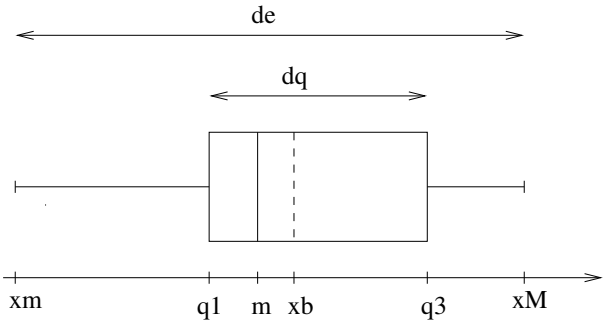


Figure 3.1 – Allure d'une boîte à moustaches.

Remarque. Il existe des variantes sur les bornes des moustaches. Plutôt que de les faire aller jusqu'à $X(1)$, $X(N)$, on peut les faire aller jusqu'aux premier et neuvième déciles, ou aux cinquième et 95-ième centiles, ou aux premier et 99-ième centiles, etc. On appelle alors outliers les données à l'extérieur des moustaches.

3.4 Paramètres de forme

3.4.1 Variable centrée réduite

Définition 3.12. Une variable X est dite *centrée* si elle est de moyenne arithmétique nulle. Elle est dite *réduite* si elle est de variance égale à 1.

Propriété 3.4. *Soit X une variable, et Y la variable définie par,*

$$Y = \frac{X - \bar{x}}{\sigma_X}.$$

Alors, la variable Y est centrée et réduite. Elle est appelée la variable centrée-réduite *associée à la variable X.*

Démonstration. Grâce aux formules de changement de variables affine pour la moyenne et la variance, on a :

$$\bar{y} = \frac{\bar{x} - \bar{x}}{\sigma_X} = 0 \quad \text{et} \quad \mathrm{Var}(Y) = \frac{1}{\mathrm{Var}(X)}\mathrm{Var}(X) = 1.$$

□

Quand on transforme une variable en la variable centrée réduite associée, on retire à cette variable toute l'information concernant son échelle ou unité, et sa localisation. Il ne reste plus que des informations sur la forme de la distribution. Cette transformation permet de comparer plusieurs variables sur le plan de la forme, même si ce sont des variables exprimées dans des échelles différentes ou qui ont des moyennes complètement différentes.

3.4.2 Moments d'une distribution

Définition 3.13.

1. On appelle *moment à l'origine* d'ordre $r \in \mathbb{N}$ le paramètre
$$m'_r = \sum_{i=1}^{p} f_i \cdot x_i^r = \frac{1}{N}\sum_{i=1}^{p} n_i \cdot x_i^r.$$
2. On appelle *moment centré* d'ordre $r \in \mathbb{N}$ le paramètre
$$m_r = \sum_{i=1}^{p} f_i (x_i - \bar{x})^r = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \bar{x})^r.$$

Remarque.

1. Pour une variable discrète, avec les données brutes, on a
$$m'_r = \frac{1}{N}\sum_{u=1}^{N} X(u)^r \quad \text{et} \quad m_r = \frac{1}{N}\sum_{u=1}^{N} (X(u) - \bar{x})^r.$$
2. La moyenne est le moment à l'origine d'ordre 1, et la variance le moment centré d'ordre 2, qui s'exprime également en fonction des moments à l'origine d'ordres 1 et 2 :
$$m_2 = \mathrm{Var}(X) = \frac{1}{N}\sum_{u=1}^{N} X(u)^2 - \bar{x}^2 = m'_2 - (m'_1)^2.$$

Chapitre 4

Statistique descriptive bi-variée

Considérons une population $\mathcal{P}$ de taille N, $\mathcal{P} = \{1, \ldots, N\}$, sur laquelle on étudie *deux* variables X et Y. Ces variables peuvent être soit qualitatives, quantitatives discrètes ou continues.

Le but de cette section est d'étudier les relations entre X et Y, ce qui revient, mathématiquement, à étudier les propriétés du *couple* (X, Y). On définit dans ce chapitre essentiellement deux types de quantités, celles dites :

- *marginales* qui ne dépendent que d'*un seul* critère mais pas des deux ;
- *conditionnelles* qui renseignent sur un critère *en fonction* des valeurs ou modalités de l'autre.

Rappelons les notations pour les séries statistiques.

- On a la *série statistique double* sou forme brute :

$$\{(X(1), Y(1)), \ldots, (X(N), Y(N))\}.$$

- Les modalités/valeurs/centre des classes de la variable X sont :

$$\{x_1, \ldots, x_p\}.$$

N.B : afin de simplifier les notations, pour une variable qualitative, on note x_i les modalités au lieu de m_i précédemment.
Les modalités/valeurs/centre des classes de la variable Y sont :

$$\{y_1, \ldots, y_q\}.$$

De plus, ces notations respectent l'ordre naturel dans le cas d'une variable qualitative ordinale ou quantitative.
Les modalités/valeurs/centre des classes du couple (X, Y) sont les $p \times q$ données :

$$\{(x_1, y_1), (x_1, y_2), \ldots, (x_p, y_q)\}.$$

Remarque.

— Désormais, l'indice i servira exclusivement à indexer des données relatives au premier membre du couple, soit X, et l'indice j exclusivement à indexer des données relatives au second, soit Y.

— Comme déjà expliqué dans le chapitre précédent, travailler avec une variable continue dont les valeurs ne sont pas regroupées en classes revient à considérer cette variable comme discrète.

4.1 Distributions marginales et conditionnelles

Il est naturel, dans un premier temps, de regarder le couple (X, Y) comme *un seul* critère d'observation. On a alors les mêmes notions qu'au chapitre 1 :

Définition 4.1.

- Soient $i \in \{1, \dots, p\}$ et $j \in \{1, \dots, q\}$. L'*effectif* du couple (x_i, y_j), noté n_{ij}, est le nombre d'individus associés à la fois à x_i pour X et à y_j pour Y et sa *fréquence*, noté f_{ij}, est le quotient de son effectif par la taille de la population :

$$n_{ij} = \text{card}\{u \in \mathcal{P} | X(u) = x_i \text{ et } Y(u) = y_j\} \quad \text{et} \quad f_{ij} = \frac{n_{ij}}{N}.$$

- Soit $i \in \{1, \dots, p\}$ (respectivement $j \in \{1, \dots, q\}$). L'*effectif marginal* de x_i (respectivement y_j), noté $n_{i\cdot}$ (respectivement $n_{\cdot j}$), est l'effectif de x_i (respectivement y_j) pour la série simple de X (respectivement Y) :

$$n_{i\cdot} = \sum_{j=1}^{q} n_{ij} \quad \text{et} \quad n_{\cdot j} = \sum_{i=1}^{p} n_{ij}.$$

- La *fréquence marginale* de x_i (respectivement y_j), notée $f_{i\cdot}$ (respectivement $f_{\cdot j}$), est alors le quotient de l'effectif marginal de x_i (respectivement y_j) par la taille de la population :

$$f_{i\cdot} = \frac{n_{i\cdot}}{N} = \sum_{j=1}^{q} f_{ij} \quad \text{et} \quad f_{\cdot j} = \frac{n_{\cdot j}}{N} = \sum_{i=1}^{p} f_{ij}.$$

Remarque. Les effectifs/fréquences marginaux sont également les effectifs/fréquences pour les séries statistiques simples de X ou de Y.

Propriété 4.1.

$$\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = \sum_{i,j} n_{ij} = N, \quad \sum_{i,j} f_{ij} = 1,$$
$$\sum_{i=1}^{p} n_{i\cdot} = \sum_{j=1}^{q} n_{\cdot j} = N \quad \textit{et} \quad \sum_{i=1}^{p} f_{i\cdot} = \sum_{j=1}^{q} f_{\cdot j} = 1.$$

Démonstration. Les deux premières égalités sont évidentes. Par symétrie entre X et Y, il suffit de démontrer le résultat pour les sommes des $n_{i\cdot}$ et $f_{i\cdot}$. De plus, en divisant par N, on déduit du résultat sur la somme des $n_{i\cdot}$ celui sur la somme des $f_{i\cdot}$. On calcule donc :

$$\sum_{i=1}^{p} n_{i\cdot} = \sum_{i=1}^{p} \Big(\sum_{j=1}^{q} n_{ij} \Big) = \sum_{i,j} n_{ij} = N,$$

comme attendu. □

Définition 4.2. Le *tableau de contingence des effectifs* du couple de variables (X, Y) est un tableau dans lequel les valeurs/classes/modalités de X sont en lignes, celles de Y en colonnes et, pour tout $i \in \{1, \dots, p\}$ et $j \in \{1, \dots, q\}$, l'effectif n_{ij} se situe à l'intersection de la ligne i

et de la colonne j . On rajoute une colonne à droite qui contient les effectifs marginaux des x_i (obtenus en faisant la somme des effectifs sur les colonnes) et une ligne en bas qui contient les effectifs marginaux des y_j (obtenus en faisant la somme des effectifs sur les lignes). La case en bas à droite contient l'effectif total N.

On peut aussi faire le *tableau de contingence des fréquences* avec les fréquences/fréquences marginales plutôt que les effectifs/effectifs marginaux.

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_q	Total
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$			$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$			$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot q}$	N

Tableau de contingence des effectifs du couple de variables (X, Y)

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_q	Total
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1q}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$			$\vdots$
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{iq}	$f_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$			$\vdots$
x_p	f_{p1}	$\cdots$	f_{pj}	$\cdots$	f_{pq}	$f_{p\cdot}$
Total	$f_{\cdot 1}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot q}$	1

Tableau de contingence des fréquences du couple (X, Y).

Remarque. Un tableau de contingence se lit comme une matrice : le premier indice est celui de la ligne et le second celui de la colonne.

En extrayant une colonne (respectivement une ligne) du tableau de contingence, on obtient une *distribution conditionnelle* de X (respectivement de Y). Précisément, l'extraction de la colonne j (respectivement de la ligne i) permet de répondre à la question « Sachant que Y vaut y_j (respectivement X vaut x_i), comment se comporte X (respectivement Y) ? » Ainsi, une distribution conditionnelle est l'étude d'un critère lorsque l'autre reste fixe – à une valeur ou modalité connue. On note $X|_{Y=y_j}$ (respectivement $Y|_{X=x_i}$) la *distribution conditionnelle de X sachant $Y = y_j$* (respectivement de Y *sachant* $X = x_i$).

Définition 4.3. Soient $i \in \{1, \ldots, p\}$ et $j \in \{1, \ldots, q\}$. La *fréquence conditionnelle* de x_i *sachant* y_j (respectivement y_j *sachant* x_i), notée $f_{i|Y=y_j}$ ou $f_{i|j}$ (respectivement $f_{j|X=x_i}$ ou $f_{j|i}$), est la fréquence d'occurrence de x_i dans la colonne j (respectivement y_j dans la ligne i) du tableau de contingence :

$$f_{i|Y=y_j} = \frac{n_{ij}}{n_{\cdot j}} \quad \text{et} \quad f_{j|X=x_i} = \frac{n_{ij}}{n_{i\cdot}}.$$

Propriété 4.2. *Les fréquences conditionnelles vérifient les égalités suivantes :*

$$\forall j \in \{1,\dots,q\},\ \sum_{i=1}^{p} f_{i|Y=y_j} = 1 \quad et \quad \forall i \in \{1,\dots,p\},\ \sum_{j=1}^{q} f_{j|X=x_i} = 1.$$

Démonstration. De nouveau par symétrie entre X et Y, il suffit de montrer, par exemple, la première égalité. Soit donc $j \in \{1,\dots,q\}$. On écrit :

$$\sum_{i=1}^{p} f_{i|Y=y_j} = \sum_{i=1}^{p} \frac{n_{ij}}{n_{\cdot j}} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} n_{ij} = \frac{1}{n_{\cdot j}} n_{\cdot j} = 1. \qquad \square$$

Définition 4.4. Le tableau des *profils-colonne* représente les distributions conditionnelles de $X|_Y$. On ajoute une ligne en bas avec les totaux, égaux à 1 d'après la propriété ci-dessus. Le tableau des *profils-ligne* représente les distributions conditionnelles $Y|_X$. On ajoute une colonne à droite avec les totaux, égaux à 1 d'après la propriété ci-dessus.

$X\|_Y$ \ Y	y_1	$\cdots$	y_q
x_1	$f_{1\|Y=y_1}$	$\cdots$	$f_{1\|Y=y_q}$
$\vdots$	$\vdots$		$\vdots$
x_p	$f_{p\|Y=y_1}$	$\cdots$	$f_{p\|Y=y_q}$
Total	1	$\cdots$	1

Tableau des profils-colonne

X \ $Y\|_X$	y_1	$\cdots$	y_q	Total
x_1	$f_{1\|X=x_1}$	$\cdots$	$f_{q\|X=x_1}$	1
$\vdots$	$\vdots$			$\vdots$
x_p	$f_{1\|X=x_p}$	$\cdots$	$f_{q\|X=x_p}$	1

Tableau des profils-ligne

Remarque. Un calcul simple donne pour tout $i \in \{1,\dots,p\}$ et $j \in \{1,\dots,q\}$:

$$f_{ij} = f_{i|Y=y_j} f_{\cdot j} \quad \text{et} \quad f_{ij} = f_{j|X=x_i} f_{i\cdot}.$$

4.2 Indépendance

Une question centrale dans les relations entre variables est celle de l'indépendance.

4.2.1 Définition

Définition 4.5. Les variables X et Y sont dites *indépendantes* lorsque :

$$\forall (i,j) \in \{1,\dots,p\} \times \{1,\dots,q\},\ f_{ij} = f_{i\cdot} f_{\cdot j}. \tag{4.1}$$

Proposition 4.1. *Soient deux variables X, Y. Les assertions suivantes sont équivalentes :*

(i) *X et Y sont indépendantes.*

(ii) $\forall (i,j) \in \{1,\dots,p\} \times \{1,\dots,q\},\ n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$.

(iii) $\forall (i,j) \in \{1,\dots,p\} \times \{1,\dots,q\},\ f_{i|Y=y_j} = f_{i\cdot}$.

(iv) $\forall (i,j) \in \{1,\dots,p\} \times \{1,\dots,q\},\ f_{j|X=x_i} = f_{\cdot j}$.

Démonstration. Il suffit de modifier l'écriture de l'égalité (4.1) définissant l'indépendance. On considère donc $i \in \{1, \dots, p\}$ et $j \in \{1, \dots, q\}$. Pour l'équivalence entre les points (i) et (ii) on écrit :

$$f_{ij} = f_{i\cdot} f_{\cdot j} \iff \frac{n_{ij}}{N} = \frac{n_{i\cdot}}{N}\frac{n_{\cdot j}}{N} \iff n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}.$$

Pour l'équivalence entre les points (ii) et (iii) on écrit :

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N} \iff \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \iff f_{i|Y=y_j} = f_{i\cdot}.$$

Et de façon similaire :

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N} \iff \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{N} \iff f_{j|X=x_i} = f_{\cdot j},$$

on obtient l'équivalence entre les points (ii) et (iv). □

4.2.2 Écart à l'indépendance

Étant donné les variables X et Y, on s'intéresse à savoir si elles sont indépendantes ou non ; et à quantifier l'écart à l'indépendance. Pour cela, on peut calculer la quantité suivante.

Définition 4.6. Le *chi-carré observé* est :

$$\chi^2 = N \sum_{i,j} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \sum_{i,j} \Big(\frac{n_{i\cdot} n_{\cdot j}}{N}\Big)^{-1} \Big(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N}\Big)^2.$$

Propriété 4.3. *Le χ^2 est toujours positif ou nul et,*

$$\chi^2 = 0 \iff \textit{les variables } X \textit{ et } Y \textit{ sont indépendantes.}$$

Il se trouve que le χ^2 observé croît avec la taille de la population. Pour cette raison, il est utile de le renormaliser de manière adéquate.

Définition 4.7. Le *coefficient V de Cramer* du couple de variables (X, Y) est :

$$V = \sqrt{\frac{\chi^2}{N \min(p-1, q-1)}}.$$

Propriété 4.4. *Le coefficient V de Cramer vérifie :*

$$0 \leq V \leq 1,$$

et $V = 0$ si et seulement si les variables sont indépendantes.

Ainsi le V de Cramer permet de mesurer l'écart à l'indépendance et sa taille ne dépend pas de la taille de la population.

4.3 Nuage de points, résumés numériques

Cette section n'a de sens que si les variables X et Y sont quantitatives, discrètes ou continues et regroupées en classes, ce que nous supposons dans la suite.

4.3.1 Nuage de points

Rappelons que nous avons a disposition la statistique double,

$$\{(X(1), Y(1)), \ldots, (X(N), Y(N))\}.$$

Cette dernière peut être représentée par un nuage de N points dans $\mathbb{R}^2$, noté $\mathcal{N}(\mathcal{P})$:

$$\mathcal{N}(\mathcal{P}) = \{(X(u), Y(u)) \in \mathbb{R}^2 : u \in \{1, \ldots, N\}\}.$$

4.3.2 Moyenne, variance, écart-type

En considérant les variables X et Y indépendamment l'une de l'autre, on peut définir tous les paramètres du chapitre 3, que l'on qualifie alors de *marginal*.

Les moyennes marginales

En reprenant les définitions du chapitre 3, nous avons :

$$\bar{x} = \sum_{i=1}^{p} f_{i\cdot} x_i = \sum_{i=1}^{p} \Big(\sum_{j=1}^{q} f_{ij} \Big) x_i = \sum_{i,j} f_{ij} x_i \quad \text{et} \quad \bar{y} = \sum_{j=1}^{q} f_{\cdot j} y_j = \sum_{j=1}^{q} \Big(\sum_{i=1}^{p} f_{ij} \Big) y_j = \sum_{i,j} f_{ij} y_j.$$

Si les variables X et Y sont discrètes, à partir des données brutes, on a :

$$\bar{x} = \frac{1}{N} \sum_{u=1}^{N} X(u), \quad \text{et} \quad \bar{y} = \frac{1}{N} \sum_{u=1}^{N} Y(u).$$

Définition 4.8. Le point $G = (\bar{x}, \bar{y})$ est appelé le *barycentre* ou *centre de gravité* du nuage de points.

La proposition suivante montre que la moyenne est linéaire.

Proposition 4.2. *Soient $a, b \in \mathbb{R}$. La moyenne $\bar{z}$ de la variable $Z = aX + bY$ s'écrit :*

$$\bar{z} = a\bar{x} + b\bar{y}.$$

Démonstration. La variable $Z = aX + bY$ prend comme valeurs $z_{ij} = ax_i + by_j$, pour $i \in \{1, \ldots, p\}$, $j \in \{1, \ldots, q\}$, avec f_{ij} comme fréquence pour la valeur z_{ij}. Ainsi, sa moyenne $\bar{z}$ est :

$$\bar{z} = \sum_{i,j} f_{ij} z_{ij} = \sum_{i,j} f_{ij}(ax_i + by_j) = a \sum_{i,j} f_{ij} x_i + b \sum_{i,j} f_{ij} y_j = a\bar{x} + b\bar{y}. \qquad \square$$

Les variances et écarts-types marginaux

D'après le chapitre 3, nous avons :

$$\mathrm{Var}(X) = \sum_{i=1}^{p} f_{i\cdot}(x_i - \bar{x})^2 = \sum_{i=1}^{p}\Big(\sum_{j=1}^{q} f_{ij}\Big)(x_i - \bar{x})^2 = \sum_{i,j} f_{ij}(x_i - \bar{x})^2$$

$$\mathrm{Var}(Y) = \sum_{j=1}^{q} f_{\cdot j}(y_j - \bar{y})^2 = \sum_{i=1}^{p}\Big(\sum_{i=1}^{p} f_{ij}\Big)(y_j - \bar{y})^2 = \sum_{i,j} f_{ij}(y_j - \bar{y})^2.$$

Les écarts-types σ_X, σ_Y sont les racines des variances $\mathrm{Var}(X), \mathrm{Var}(Y)$.

Si les variables X et Y sont discrètes, à partir des données brutes, on a :

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{u=1}^{N}(X(u) - \bar{x})^2, \quad \text{et} \quad \mathrm{Var}(Y) = \frac{1}{N}\sum_{u=1}^{N}(Y(u) - \bar{y})^2.$$

Attention, à la différence de la moyenne, la variance n'est pas linéaire. Afin de calculer la variance d'une combinaison linéaire de variables, nous devons d'abord introduire la covariance.

4.3.3 Covariance

Soient X, Y deux variables quantitatives.

Définition 4.9. La *covariance* de X et Y, notée $\mathrm{Cov}(X, Y)$, est la moyenne des produits des écarts des valeurs ou centres des classes de X et Y à leur moyenne respective :

$$\mathrm{Cov}(X, Y) = \sum_{i,j} f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{N}\sum_{i,j} n_{ij}(x_i - \bar{x})(y_j - \bar{y}).$$

Remarque.

1. La covariance possède une unité : celle du produit des unités de X et Y.
2. La covariance est symétrique : $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
3. En faisant $Y = X$, on obtient $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
4. Si les variables X et Y sont discrètes, à partir des données brutes on a :

$$\mathrm{Cov}(X, Y) = \frac{1}{N}\sum_{u=1}^{N}(X(u) - \bar{x})(Y(u) - \bar{y}).$$

Proposition 4.3. *La covariance de X et Y vérifie les propriétés suivantes :*

1. La covariance de X et Y s'exprime également comme :

$$\mathrm{Cov}(X, Y) = \Big(\sum_{i,j} f_{ij} x_i y_j\Big) - \bar{x}\,\bar{y} = \Big(\frac{1}{N}\sum_{i,j} n_{ij} x_i y_j\Big) - \bar{x}\,\bar{y}.$$

Si les variables X et Y sont discrètes, à partir des données brutes on a :

$$\mathrm{Cov}(X, Y) = \Big(\frac{1}{N}\sum_{u=1}^{N} X(u)Y(u)\Big) - \bar{x}\bar{y}.$$

2. *Soient $a, b, c, d \in \mathbb{R}$. On a la formule de changement de variables suivante :*
$$\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y).$$
3. *Soient $a, b \in \mathbb{R}$. La variance de la variable $aX + bY$ s'écrit :*
$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X, Y) + b^2\,\mathrm{Var}(Y).$$

Démonstration. Pour le point 1, on développe la formule de la covariance :
$$\begin{aligned}\mathrm{Cov}(X, Y) &= \sum_{i,j} f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sum_{i,j} f_{ij}(x_i y_j - x_i\bar{y} - y_j\bar{x} + \bar{x}\,\bar{y}) \\ &= \Big(\sum_{i,j} f_{ij}x_i y_j\Big) - \bar{y}\Big(\sum_{i,j} f_{ij}x_i\Big) - \bar{x}\Big(\sum_{i,j} f_{ij}y_j\Big) + \bar{x}\,\bar{y}\Big(\sum_{i,j} f_{ij}\Big),\end{aligned}$$
et en utilisant :
$$\bar{x} = \sum_{i,j} f_{ij}x_i, \quad \bar{y} = \sum_{i,j} f_{ij}y_j \quad \text{et} \quad \sum_{i,j} f_{ij} = 1,$$
on obtient le résultat.

Pour le point 2, on reprend la définition de la covariance sachant que les moyennes de $aX + b$ et $cY + d$ sont respectivement $a\bar{x} + b$ et $c\bar{y} + d$:
$$\begin{aligned}\mathrm{Cov}(aX + b, cY + d) &= \sum_{i,j} f_{ij}\big((ax_i + b) - (a\bar{x} + b)\big)\big((cy_j + d) - (c\bar{y} + d)\big) \\ &= ac\sum_{i,j} f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = ac\,\mathrm{Cov}(X, Y).\end{aligned}$$

Pour le point 3, posons $Z = aX + bY$. Alors, d'après la proposition 4.2, $\bar{z} = a\bar{x} + b\bar{y}$. Ainsi,
$$\begin{aligned}\mathrm{Var}(aX + bY) = \mathrm{Var}\,Z &= \sum_{i,j} f_{ij}(z_{ij} - \bar{z})^2 \\ &= \sum_{i,j} f_{ij}\big((ax_i + by_j) - (a\bar{x} + b\bar{y})\big)^2 \\ &= \sum_{i,j} f_{ij}\big(a(x_i - \bar{x}) + b(y_j - \bar{y})\big)^2 \\ &= a^2\sum_{i,j} f_{ij}(x_i - \bar{x})^2 + 2ab\sum_{i,j} f_{ij}(x_i - \bar{x})(y_j - \bar{y}) + b^2\sum_{i,j} f_{ij}(y_j - \bar{y})^2 \\ &= a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X, Y) + b^2\,\mathrm{Var}(Y),\end{aligned}$$
□

La covariance permet de quantifier la dépendance de X et Y dans le sens suivant :

Théorème 4.1. *Si les variables X et Y sont indépendantes, alors* $\mathrm{Cov}(X, Y) = 0$.

Démonstration. On reprend l'écriture de la covariance du point 1. de la proposition 4.3, que l'on combine avec les expressions des moyennes marginales :
$$\begin{aligned}\mathrm{Cov}(X, Y) &= \Big(\sum_{i,j} f_{ij}x_i y_j\Big) - \bar{x}\,\bar{y} = \Big(\sum_{i,j} f_{ij}x_i y_j\Big) - \Big(\sum_{i=1}^{p} f_{i\cdot}x_i\Big)\Big(\sum_{j=1}^{q} f_{\cdot j}y_j\Big) \\ &= \Big(\sum_{i,j} f_{ij}x_i y_j\Big) - \Big(\sum_{i,j} f_{i\cdot}f_{\cdot j}x_i y_j\Big) = \sum_{i,j}(f_{ij} - f_{i\cdot}f_{\cdot j})x_i y_j.\end{aligned}$$

Or, par définition, si X et Y sont indépendantes, pour tout $i \in \{1, \dots, p\}$, $j \in \{1, \dots, q\}$, on a $f_{ij} = f_{i\cdot} f_{\cdot j}$; d'où $\mathrm{Cov}(X, Y) = 0$. □

Corollaire 4.2. *Si les variables X et Y sont indépendantes, alors :*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

4.3.4 Coefficient de corrélation

Théorème 4.3 (Inégalité de Cauchy-Schwarz)**.** *Soient X, Y des variables discrètes ou continue regroupées en classes. On a l'inégalité suivante :*

$$[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y) \quad \textit{ou de façon équivalente} \quad |\,\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y. \tag{4.2}$$

De plus, si les variables X et Y ne sont pas constantes, alors l'égalité dans l'équation (4.2) *a lieu si et seulement si :*

$$Y = \epsilon \frac{\sigma_Y}{\sigma_X} X + b \quad \textit{avec} \quad \epsilon \in \{-1, 1\} \textit{ et } b \in \mathbb{R}.$$

Démonstration. On considère la variable $\lambda X + Y$ où $\lambda \in \mathbb{R}$. On a donc d'après le point 3. de la proposition 4.3 :

$$\mathrm{Var}(\lambda X + Y) = \lambda^2 \,\mathrm{Var}(X) + \lambda\big(2\,\mathrm{Cov}(X, Y)\big) + \mathrm{Var}(Y),$$

que l'on voit comme un trinôme du second degré en λ de discriminant :

$$[2\,\mathrm{Cov}(X, Y)]^2 - 4\,\mathrm{Var}(X)\,\mathrm{Var}(Y) = 4[\mathrm{Cov}(X, Y)]^2 - 4\,\mathrm{Var}(X)\,\mathrm{Var}(Y).$$

De plus, comme $\mathrm{Var}(\lambda X + Y) \geq 0$, on sait que ce discriminant est négatif, d'où :

$$\begin{aligned}[\mathrm{Cov}(X, Y)]^2 - \mathrm{Var}(X)\,\mathrm{Var}(Y) \leq 0 &\iff [\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y) \\ &\iff |\,\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y,\end{aligned}$$

ce qui montre la première partie. De plus, l'égalité dans l'équation (4.2) signifie d'une part que $\mathrm{Cov}(X, Y) = \epsilon \sigma_X \sigma_Y$ pour un certain $\epsilon \in \{-1, 1\}$ et d'autre part que le discriminant est nul, donc que le polynôme admet une (unique, car X et Y ne sont pas constantes) racine :

$$\lambda_0 = -\frac{2\,\mathrm{Cov}(X, Y)}{2\,\mathrm{Var}(X)} = -\epsilon \frac{\sigma_Y}{\sigma_X}.$$

Mais alors, la variable $\lambda_0 X + Y$ est de variance nulle, donc constante de valeur notée $b \in \mathbb{R}$, d'après le théorème 3.1. □

Définition 4.10. Lorsque les variables X et Y ne sont pas constantes, le *coefficient de corrélation* de X et Y, noté $r(X, Y)$, est le quotient de la covariance de X et Y par le produit des écarts-type de X et Y :

$$r(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Corollaire 4.4 (du théorème 4.3)**.** *Le coefficient de corrélation de X et Y vérifie :*

$$-1 \leq r(X,Y) \leq 1.$$

De plus, on a $|r(X,Y)| = 1$ si et seulement si X et Y vérifient une relation affine de la forme $Y = aX + b$, avec $b \in \mathbb{R}$ et :

$$r(X,Y) = 1 \iff a = \frac{\sigma_Y}{\sigma_X} > 0 \quad et \quad r(X,Y) = -1 \iff a = -\frac{\sigma_Y}{\sigma_X} < 0.$$

Remarque. Lorsque le coefficient de corrélation est « proche » de 1 en valeur absolue – prenons $|r(X,Y)| \geq 0,9$ pour fixer les idées –, cela signifie qu'il y a une *liaison affine* importante entre les variables X et Y ; cette liaison étant exacte lorsque $|r(X,Y)| = 1$. En revanche, lorsque le coefficient de corrélation est proche de 0, alors il n'existe pas de relation affine entre X et Y, ce qui ne signifie pas qu'un autre type de liaison ne peut avoir lieu.

4.4 Ajustement affine

On suppose ici qu'il y a une liaison affine importante entre X et Y. Le principe d'un *ajustement affine*, aussi appelé *régression linéaire*, est de déterminer la « meilleure » droite pour exprimer la liaison affine entre X et Y, c'est-à-dire la droite qui approxime au mieux le nuage de points $\mathcal{N}(\mathcal{P})$.

Remarque. Dans les cas où la liaison entre X et Y est « indirectement » affine, c'est-à-dire de la forme $Y = af(X) + b$, on détermine dans un premier temps la fonction f. On étudie ensuite la variable $Z = f(X)$ et l'on effectue une régression linéaire sur le couple (Z,Y). C'est une *régression linéaire après changement de variable.*

4.4.1 Méthode des moindres carrés

Pour tous $a, b \in \mathbb{R}$, on considère la variable $\Delta_{(a,b)} = Y - (aX + b)$. Cette variable contient l'information sur l'erreur commise si l'on modélise Y à partir de X par le changement de variable affine $aX + b$. On calcule la moyenne et la variance de $\Delta_{(a,b)}$:

$$\begin{aligned}
\overline{\Delta_{(a,b)}} &= \bar{y} - (a\bar{x} + b) \quad \text{et} \\
\operatorname{Var}(\Delta_{(a,b)}) &= \operatorname{Var}\big(Y - (aX + b)\big) \\
&= \operatorname{Var}(Y) - 2\operatorname{Cov}(Y, aX + b) + \operatorname{Var}(aX + b) \\
&= \operatorname{Var}(Y) - 2a\operatorname{Cov}(X,Y) + a^2 \operatorname{Var}(X).
\end{aligned}$$

La *méthode des moindres carrés* consiste à déterminer $a, b \in \mathbb{R}$ de telle sorte que $\Delta_{(a,b)}$ soit le plus proche de la variable nulle, ce qui revient à minimiser $|\overline{\Delta_{(a,b)}}|$ et $\operatorname{Var}(\Delta_{(a,b)})$.

Proposition 4.4 (Méthode des moindres carrés)**.** *Si X n'est pas constante, la* droite de régression par la méthode des moindres carrés *de Y en fonction de X est la droite d'équation $y = \hat{a}x + \hat{b}$ dans le repère du nuage de points de (X,Y) avec :*

$$\hat{a} = r(X,Y)\frac{\sigma_Y}{\sigma_X} \quad et \quad \hat{b} = \bar{y} - r(X,Y)\frac{\sigma_Y}{\sigma_X}\bar{x}.$$